FEATURE SELECTION IN PATHOLOGICAL VOICE CLASSIFICATION USING DINAMYC OF COMPONENT ANALYSIS

M. Sarria-Paja[1], G. Daza-Santacoloma[1], J. I. Godino-Llorente[2], G. Castellanos-Domínguez[1], N. Sáenz-Lechón[2]
[1]Control and Digital Signal Processing Group, Universidad Nacional de Colombia, Colombia
[2]Dpt. of Circuits & Systems Engineering, Universidad Politécnica de Madrid, Spain

ABSTRACT

This paper presents a methodology for the reduction of the training space based on the analysis of the variation of the linear components of the acoustic features. The methodology is applied to the automatic detection of voice disorders by means of stochastic dynamic models. The acoustic features used to model the speech are: MFCC, HNR, GNE, NNE and the energy envelopes. The feature extraction is carried out by means of PCA, and classification is done using discrete and continuous HMMs. The results showed a direct relationship between the principal directions (feature weights) and the classification performance. The dynamic feature analysis by means of PCA reduces the dimension of the original feature space while the topological complexity of the dynamic classifier remains unchanged. The experiments were tested with Kay Elemetrics (DB1) and UPM (DB2) databases. Results showed 91% of accuracy with 30% of computational cost reduction for DB1.

1. INTRODUCTION

During phonation of sustained vowels, the normal voice is a regular and periodic signal; however changes in its waveform and power spectrum can be appreciated if some disorders arise. For instance, the disorders in vocal folds dynamics produce a turbulent flow through the glottis affecting the voice with high-frequency noise, alterations in the formants' structure and harmonic components [1]. Therefore, the spectral distribution, the energy, and the fundamental frequency envelopes, which are not computationally intensive to measure, can be reliably employed for large-scale, rapid assessment of normal and pathological voices [2]. Moreover, the classical distortion measures based on fluctuations of acoustic measures may be complemented with dynamic features obtained from its contours, as pointed out by other studies [3]. As an illustration, it is demonstrated the potential value of the windowed relative deviation spectrum for the diagnosis, which reveals the time locations and extent of the pitch transitions, intonations, and more subtle structures that may be related to human voice control mechanisms [4].

One of the key properties that make dynamic features useful is that they consider changes in the temporal structure of the excitation signal. Mainly, static classifiers remove all temporal dependency and therefore dynamic pattern classifiers are needed to handle explicit temporal dependencies in the pathological voices [5]. In this regard, diverse approaches have been proposed. In [6], it is presented an insight of the effects of the application of autoregressive decomposition and pole tracking to pathological voice signals. Recently, the application of non-linear dynamical techniques to normal and pathological voices has become a strong focus of research activity, because they place more realistic assumptions on the voice production mechanism [7, 8]. Nonetheless, pathological voices involve more irregularities, such as hoarseness, breathiness, and vocal instability; thus, the applicability of nonlinear dynamics to pathological speech data may require large amounts of data.

Lastly, short term features combined with dynamic classifiers (e.g. Hidden Markov Models - HMM), have been used in the classification of pathological voices [5]. But a significant limitation of the standard HMM is the way it models the state durations. Besides, it is no clear whether gathering of dynamic features should lead to an improved representation capability, and hence to higher performance of the dynamic classifier. Therefore, a more detailed study should be conducted to assess the relevance of dynamic features that describe pathologies, which could be used in data analysis and evaluation to support diagnose by automatic dynamic classifiers [9].

The aim of this paper is the comparison and evaluation of dynamic feature sets that are suitable for classification of pathological voices using HMM. For this purpose, the feature selection methodology presented in [10], is adopted. This method uses Principal Component Analysis (PCA) to evaluate temporal relations on the set of dynamic features and project them onto lower dimensional subspaces. The relevance is obtained by examining the principal directions.

2. RELEVANCE OF DYNAMIC FEATURES

Widely known approaches, like PCA [11, 12, 13] and sequential search methods, have been customized as feature selection methods for the use with a HMM classifier. Assuming that the input contour data are highly correlated, linear transformation methods such as PCA tries to exploit the correlation present in the data by projecting the data onto a new space where the axes are orthogonal to each other.

Let $\xi_{ij}[k], k=1,\ldots,m$ be the *j*-th dynamic feature belonging to *i*-th observation, where $j=1,\ldots,p$, $i=1,\ldots,n$; being n the number of observations and p the number of features, which change over time *k*. Each vector observation $\boldsymbol{\xi}_i$ can be represented by a supervector of size $mp\times 1$:

$$\boldsymbol{\xi}_i = \begin{matrix}[\xi_{i1}[1], \xi_{i1}[2], \cdots, \xi_{i1}[m], \xi_{i2}[1], \cdots \\ \cdots, \xi_{i2}[m], \cdots, \xi_{ip}[m]]^{\mathrm{T}}\end{matrix} \quad (1)$$

The respective covariance matrix, after centering each one of the observation supervectors is computed as:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i^0\boldsymbol{\xi}_i^{0T} = \frac{1}{n}\mathbf{G}\mathbf{G}^{\mathrm{T}} \quad (2)$$

Where $\mathbf{G}$ stands for matrix $\mathbf{G}=\left[\boldsymbol{\xi}_1^0\ \boldsymbol{\xi}_2^0\cdots\boldsymbol{\xi}_n^0\right]$. In most cases, we are far away from computing the eigenvectors $\mathbf{v}$ and eigenvalues $\boldsymbol{\lambda}$ of such a huge matrix. Nevertheless, the rank properties of $\mathbf{G}$ can be used, in special, the one that state that $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ has the same non-null eigenvalues than $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ and the advantage of $n \ll pm$, as given in [14]:

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\hat{\mathbf{v}}_i = \lambda\hat{\mathbf{v}}_i \quad (3)$$

being $\hat{\mathbf{v}}_i$ the eigenvectors of $\mathbf{G}^{\mathrm{T}}\mathbf{G}$, so that, $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i$. Therefore, the eigenvectors corresponding to non-zero eigenvalues of $\mathbf{S}$ are $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i / \left\|\mathbf{G}\hat{\mathbf{v}}_i\right\|$. The eigenvectors associated with the r largest eigenvalues of $\mathbf{S}$ are selected as Principal Directions [15], which span an orthonormal basis for a subspace containing most of the information given by observations. Trying to reproduce the observation in the original space as a linear combination of the r principal directions,

$$\hat{\boldsymbol{\xi}}_i^0 = \sum_{k=1}^{r} w_k \mathbf{v}_k^T \quad (4)$$

so, from (4) the reconstruction weights $w_k = \mathbf{v}_k^T\boldsymbol{\xi}_i^0$ can be though as the new set of features, and taking advantage of the orthonormality property of the basis, observations can be recognized using geometric criteria to partition the subspace off.

On the other hand, the proposed method allows for identifying and choosing those dynamic features that influence the most. The magnitudes of the entries of the eigenvectors that span the representation basis, tell us the variables to be choose. Let $\boldsymbol{\rho}$ be the vector expressed as; $\boldsymbol{\rho} = \sum_{k=1}^{r}\left|\lambda_k\mathbf{v}_k\right|$, so, that its larger values are the most significant windows from the dynamic features. Rearranging $\boldsymbol{\rho}$ in the following manner:

$$\boldsymbol{\rho} = \left[\rho_{11}\ \rho_{12}\ \cdots\rho_{1m}\ \rho_{21}\ \cdots\rho_{2m}\ \cdots\rho_{p1}\ \cdots\rho_{pm}\right]^{\mathrm{T}} \quad (5)$$

$$\Rightarrow \mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{p2} \\ \vdots & \vdots & & \vdots \\ \rho_{1m} & \rho_{2m} & \cdots & \rho_{pm} \end{bmatrix}$$

it is possible to obtain the scalar $\hat{\rho}_j = \sum_{k=1}^{m}\rho_{jk}$, $j=1,\ldots,p$ which is the sum of the elements of each column *j* from $\mathbf{P}$ matrix. In consequence, the main assumption is that the largest values of $\hat{\rho}_j$ point out to the best input attributes since they exhibit higher overall correlations with principal components.

3. EXPERIMENTAL SETUP

3.1. Databases

Kay Elemetrics (DB1) and UPM (DB2) databases of voice disorders (described in [9,10]) were used to test the proposed methodology. From DB1 a set of 173 pathological and 53 normal speakers has been taken, the recorded material is the sustained phonation of /ah/ vowel from patients with a variety of voice pathologies: organic, neurological, and traumatic disorders [11, 12]. The DB2 stores 239 pathological voices with a wide variety of organic pathologies (nodules, polyps, edemas, carcinomas, etc), and 201 normal voices. The dataset contains the sustained phonation of the /a/ Spanish vowel with a sampling rate of 50 kHz and 16-bits of resolution. Each database was split in two disjoint subsets (training and test sets), and each recorded voice (observation) was uniformly windowed employing 40 ms length window with 50% of overlapping. Within each window 48 features were computed. These features correspond to 16 measures and its first and second derivatives. These measures are: 12 Mel Frequency Cepstrum Coefficients (MFCC)[16], the Harmonics to Noise Ratio (HNR)[17], the Glottal to Noise Excitation Ratio (GNE)[18], the Normalized Noise Energy (NNE) [19], and the Energy of the frame, as well.

The accuracy was measured using a *k-folds* cross validation strategy. In particular, 11-folds have been used, splitting the 70% of the files for training the classifier, and the remaining 30% for validating. These sets were randomly chosen.

3.2. HMM training

For each observation are taken *j* (*j* = 1, 2, …, 48) dynamic features. These features were selected using the relevance measure presented in section 2. Firstly, we worked with discrete HMMs. We create a codebook with *K* symbols (*K*=64 or *K*=128) and we used the Linde-Buzo-Gray (LBG) algorithm to generate a quantized version of the training set. This information is employed for training the models for each class using Baum-Welch algorithm.

The *j* first contours of each test observation (not used yet) are quantized by means of LBG, so the observation symbols are taken from the codebook generated in the training stage. To determine which class a register belongs to (pathological or normal voice) the Maximum A-Posterior (MAP) rule were employed.

Secondly, we used continuous HMMs. With this classification approach, it is necessary to guarantee an adequate initial parameter set for estimating the density function of the observations (i.e. mean vectors and covariance matrix). An accurate estimation of the initial parameters is essential for a quick and proper convergence of the reestimation formulas. This procedure is carried out by means of segmental k-means. After initial parameter estimation we iteratively estimate the class models by means of the Baum-Welch algorithm. As a result of the relevance analysis carried out above, a set of weights for the features was obtained. Fig. 1 shows the weights for each one of the 48 features.

The feature set used in the classifier was incremented according to the feature weights (in descending order, from the best up to the worst feature) to improve the classification accuracy (Fig. 2(c)). Figures 2(a) and 2(b) show the classification accuracy when the features in the classifier are incremented randomly and in ascending order (from the worst up to the most significant feature) according to the feature weights, respectively.

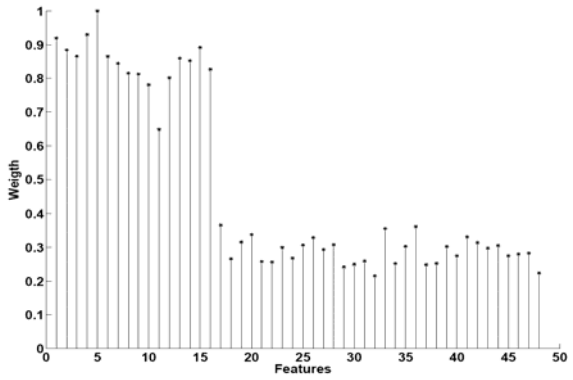

Fig. 1 Relevance of dynamic features – DB1

4. RESULTS AND DISCUSSION

4.1. Discrete observation density

Initially, the experiments were carried out with discrete HMMs using DB1. Features used by the classifier were incremented while the accuracy was measured. The number of states of the HMM tested in the experiments was chosen between 3 and 10 (NS=3, NS=5, and NS=10), and the codebook dimensions between 64 and 128 (CB=64, and CB=128).

In a first approximation, the incremental set of features was randomly chosen. In other words, the feature weights were no considered. From Fig. 2 it is possible to observe that the maximum performance is reached with a small number of features, this means that it is not necessary to use the whole feature set. Nonetheless, since the incremental set was not appropriately chosen, we can see strong oscillations in the accuracy while the feature set is incremented.

For randomly chosen features, the increase of the feature set does not present a monotonous behavior improving the performance: the performance increases or decreases arbitrarily. This is because when we add a relevant feature the classification accuracy can abruptly grow, while when we add non-relevant features the improvement of performance may be not notorious, indeed the performance may worsen. Besides, the best accuracy is obtained with the full feature set.

Secondly, we incremented the feature set in ascending order according to the weights calculated (from the least relevant to the more relevant feature). The classification results are shown in Fig. 2(b). In this case, graphic exposes an increase monotonous accuracy behavior. However, a good performance is just reached with a large number of features (45 features). Particularly, around 90% the accuracy obtained is independent of both the codebook dimension and the number of states.

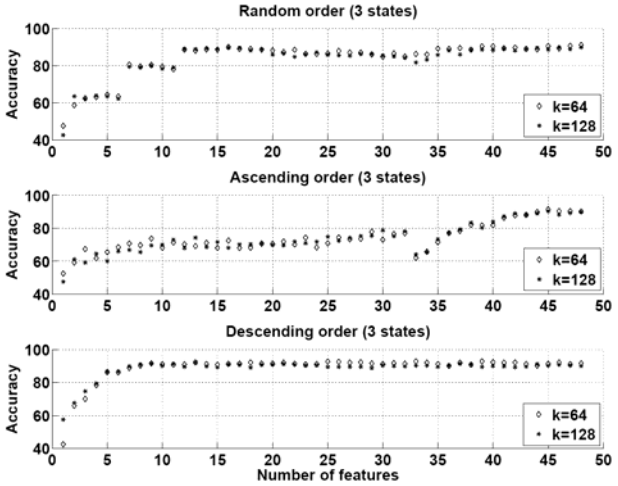

Fig. 2. Accuracy (in %) obtained according to set of features is increased. (a) *Top:* Randomly, (b) *Center:* Weights in ascending order (c) *Bottom:* Weights in descending order.

Finally, the most interesting results were achieved when the feature set is incremented according to the relevance in descending order (Fig. 2(c)). It is clear from Fig. 2(c) that the best performance was reached with a small number of features (around 90%). After achieving this performance the accuracy is almost constant. Therefore it is not necessary to use the full feature set. The results showed that a small number of features is enough. Besides, adding features does not strongly worse the performance. Moreover, as we used only a reduced feature subset then the computational cost was clearly reduced.

The results shown in Fig 1 are the starting point for selection and subsequent reduction of features, this analysis combined with the results obtained (Fig 2. and Table 1) showed that the most significant features are the instantaneous measures, without the derivatives, since the most weighted are the first 16.

Table 1 shows the results obtained for DB1. When the codebook (CB) has 64 symbols the best accuracy was reached for 18 features (NF=18). Moreover we can note that the difference between the accuracy achieved with the full feature set and with the reduced feature sets (13 and 25 features) is not significant.

When the codebook has 128 symbols the best results were obtained using 13 features. In all experiments for DB1, when we use a reduced feature set, the accuracy is higher than the one for the full feature set. Furthermore, the performance is always higher than 90% when we use 9 or more features.

CB	Number of states (NS)					
	3		5		10	
	Accuracy	NF	Accuracy	NF	Accuracy	NF
64	92.1±3.0	13	91.8±2.6	13	91.5±2.9	13
	92.6±3.8	25	92.4±3.7	25	92.1±3.5	18
	91.5±2.7	48	91.1±2.8	48	90.9±3.0	48
128	92.1±3.7	13	92.1±3.5	13	91.8±3.4	13
	89.3±2.4	25	89.2±2.2	25	89.1±4.0	18
	90.0±4.0	48	90.0±4.0	48	89.9±3.5	48

Table 1. Accuracy results (in %)for DB1 using discrete HMMs.

In the case of DB2, the results are presented in Fig. 3 and in Table 2. Particularly, the incremental feature set used to classify is sorted in descending relevance order. The behavior is similar to that presented for DB1.

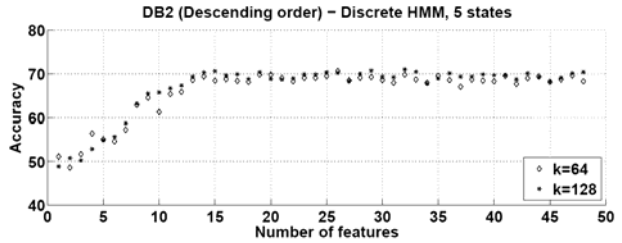

Fig 3: Accuracy (in %) as a function of the number of features for DB2 using discrete HMMs

Despite of the fact that the accuracy achieved for DB2 is lower than the accuracy obtained for DB1, the methodology showed to be consistent and it can be adequately applied to reduce the feature space.

CB	Number of states (NS)					
	3		5		10	
	Accuracy	NF	Accuracy	NF	Accuracy	NF
64	71.2±2.2	26	70.6±3.1	26	69.8±1.8	29
128	71.4±2.5	26	70.9±2.9	32	70.9±3.8	29

Table 2. best Accuracy results (in %) for DB2 using discrete HMMs.

The lower performance obtained with DB2 may be due to the diversity in the pathological class. This database has a larger number of pathologies, hence the variability of the classes is higher, and perhaps the evaluated features are not enough to model it correctly.

The best results achieved with DB2 were employing 26 features for both 64 and 128 symbols. In the first row of the Table 2 it is presented the accuracy for 15 features, and in the last row is presented the accuracy for the full feature set. It is possible to notice that there is not a significant difference among these results.

4.2. Continuous observation density

Similar to the above described experiments, the accuracy results are recalculated employing continuous HMMs. Several values of mixtures (NG=2, NG=3, and NG=4) and states (NS=3, NS=5, and NS=10) were tested. Since in the discrete case we obtained a good accuracy with a small number of features (9 up to 13), the set of features for the continuous HMM was incremented up to 32 features in descending order. Results are shown in Fig. 4 and Table 3. The best classification accuracy for continuous HMMs is reached using NG=2 and 3 states.

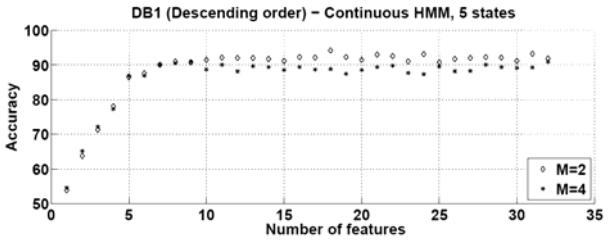

Fig. 4: Accuracy (in %) as a function of the number of features for DB1 using continuous HMMs.

NG	Number of states (NS)					
	3		5		10	
	Accuracy	NF	Accuracy	NF	Accuracy	NF
2	94.2±3.4	21	94.2±2.8	18	91.0±4.3	26
3	92.9±1.8	14	92.1±3.8	13	88.1±2.8	11
4	92.5±1.8	10	90.9±3.5	9	89.3±3.0	8

Table 3. best Accuracy results (in %) for DB1 using continuous HMMs.

4.3. Time needed to generate the codebook.

Using the same system the machine time spent to construct the codebook while the number of features is increased in descending order according to their relevance is measured. For DB1 and 64 symbols, there is a computational cost reduction of 31.05% when we use 25 features in contrast to use the full feature set. For DB2 reduction is of 26.87% between the 26 feature set and full feature set. Moreover, time needed to generate the codebook for DB1 is 87.68% lower than the time needed by DB2.

5. CONCLUSIONS

The proposed methodology for reducing the number of dynamic features in the identification of pathological voices proved to be useful for the experiments carried out. As a result was obtained an adequate performance while employing a considerably reduced feature set. The presented way of training shows that for the automatic detection of pathological voices is better to use a good set of features than a complex stochastic dynamic training model, because the later may have lower generalization capabilities.

From the incremental training, it is possible to notice that if the initial set of features is not suitable, then the model can not discriminate correctly and hence the performance decrease. But if we use a correct feature set that reflects the stochastic dynamic of the process, the performance can be augmented.

Reduction in the computational cost is clear. The results showed that the time needed to generate the codebook is highly related to the number of features employed, because the reduction of the feature set diminish the machine time spent. In general, the whole reduction of the computational cost was done employing a simple HMM architecture, which is discrete HMM.

6. ACKNOWLEDGEMENTS

This work was carried out under grants: TEC2006-12887-C02 from the Ministry of Science and Technology of Spain; 20201004208 funded by Universidad Nacional de Colombia; and *"Detección de los niveles de compromiso de resonancia en niños con labio y/o paladar hendido"* financed by DIMA.

7. REFERENCES

[1] Petar Mitev and Stefan Hadjitodorov, "Fundamental frequency estimation of voice of patients with laryngeal disorders", Information Sciences: An international journal, Volume 156, Issue 1-2 (November 2003).

[2] Alireza A. Dibazar, S. Narayanan, T. W. Berger, "Feature Analysis for Automatic Detection of Pathological Speech", Proceedings of the Second Joint EMBS/BMES conference, Houston, Texas, October 23-26, 2002.

[3] P. Gómez, J. I. Godino, F. Rodríguez, F. Díaz, V. Nieto, A. Álvarez, V. Rodellar, "Evidence of Vocal Cord Pathology From the Mucosal Wave Cepstral Contents", Acoustics, Speech, and Signal Processing, vol 5, pp 437 – 440, May 2004.

[4] J. Brandon Laflen, Ph.D., Cathy L. Lazarus, Ph.D., and Milan R. Amin, M.D. "Using Windowed Relative Deviation to Detect Possible Voice Pathology ", Proceedings of the 28th IEEE EMBS Annual International Conference, New York, August 30-Sept 3, 2006, Pages 3755 – 3758

[5] Alireza A. Dibazar, Theodore W. Berger, and Shrikanth S. Narayanan, "Pathological Voice Assessment," Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, August. 2006, pp. 1669 – 1673.

[6] Paulo R Scalassara, Maria E. Dajer and Carlos D. Maciel, "Application of Autoregressive Decomposition and Pole Tracking to Pathological Voice Signals," Proceedings of the Seventh IEEE International Symposium on Multimedia, Pages 733-738, Irvine, CA, USA. 2005

[7] C. Manfredi "Voice models and analysis for biomedical applications," Biomedical Signal Processing and Control, Volume 1, Issue 2, April 2006, Pages 99-101

[8] J. J. Jiang and Yu Zhang, "Nonlinear dynamic analysis of speech from pathological subjects," ELECTRONICS LETTERS 14th March 2002 Vol. 38 No. 6.

[9] Nicolás Sáenz-Lechón, Juan I. Godino-Llorente, Víctor Osma-Ruiz and Pedro Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," Biomedical Signal Processing and Control, Volume 1, Issue 2, April 2006, Pages 120-128.

[10] Genaro Daza-Santacoloma, Julián D. Arias-Londoño, Juan I. Godino-Llorente, Nicolás Sáenz-Lechón, Victor Osma-Ruiz, and Germán Castellanos-Domínguez, "Dynamic Feature Extraction: An Application to Voice Pathology Detection", Intelligent Automation and Soft Computing, (IN PRESS).

[11] L. Rankinea, M. Mesbaha and B. Boashash. "IF estimation for multicomponent signals using image processing techniques in the time–frequency domain", Signal Processing, vol. 87, pp. 1234-1250, 2007

[12] G. Stemmer, C. Hacker, E. Noth and H. Niemann, "Multiple Time Resolutions for Derivative s of Mel-Frequency Cepstral Coefficients", Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on, pp. 37-40, 2002. ,

[13] C. Navin-Gupta, R. Palaniappan, S. Rajan, S. Swaminathan and S.M. Krishnan, "Segmentation and Classification of Heart Sounds", CCECE/CCGEI, IEEE, pp. 1674-1677, 2005.

[14] M.Turk and A.Pentland, "Eigenfaces for recognition," Cognitive Neuroscience, vol. 3, no. 1, pp.71-86, 1991

[15] I.Jolliffe, Principal Components Analysis, Second ed., Springer, 2002.

[16] X. Huang, A. Acero and H. W. Hon, Spoken Language Processing, Prentice Hall, 2001.

[17] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," J. of Speech and Hearing Res., 36(2), pp. 254-266, 1993.

[18] D. Michaelis, T. Gramss and H. W. Strube, "Glottal-to-Noise Excitation ratio - a new measure for describing pathological voices," Acustica/Acta acustica, 83, pp. 700-706, 1997.

[19] H. Kasuya, S. Ogawa, K. Mashima and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," J. Acoust. Soc. of America, 80(5), pp. 1329-1334, 1986.